CLAIMS

1. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first tree, wherein the first set of data includes selected data of the first tree, the selected data at least partly specifying tree data;

accessing at least a second set of data of a second tree;

determining an edit sequence between at least part of the first set of data and at least part of the second set of data, the edit sequence including any of insertions, deletions, substitutions, matches, and repetitions; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the edit sequence.

2. The method of claim 1, wherein the repetitions are subtrees of at least the first tree.

3. The method of claim 1, wherein the edit sequence includes at least two repetitions, the at least two repetitions based on at least one subtree of the first tree, and the at least two repetitions appears in the second tree, and the at least two repetitions include at least a first repetition and a second repetition, and the first repetition has at least one difference from the second repetition.

4. The method of claim 3, wherein each of the at least two repetitions is obtainable from the at least one subtree of the first tree by some sequence of one or more insertions, deletions, substitutions and matches.

5. The method of claim 1, wherein the edit sequence includes none of insertions, deletions, substitutions, matches, and repetitions.

6. The method of claim 1, wherein the edit sequence includes at least one of one or more insertions of nodes, one or more insertions of subtrees, one or more deletions of subtrees, one or more deletions of subtrees, one or more substitutions of nodes, one or more substitutions of subtrees, one or more repetitions of nodes, and one or more repetitions of subtrees.

7. The method of claim 1, wherein the edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

8. The method of claim 7, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some selected data of the first tree and at least some data from the second tree.

9. The method of claim 7, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more repetitions.

10. The method of claim 7, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

11. The method of claim 7, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

12. The method of claim 7, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

13. The method of claim 7, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

14. The method of claim 7, wherein a first cost is associated with a first repetition at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second repetition at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

15. The method of claim 7, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

16. The method of claim 7, wherein data includes at least a first type and a second type, and the one or more costs are at least partly set to discourage substitutions of one or more of the first type for one or more of the second type.

17. The method of claim 7, wherein data includes at least a first type and a second type, and the one or more costs are at least partly set to discourage substitutions of one or more of the second type for one or more of the first type.

18. The method of claim 7, wherein a first cost is associated with preserving data of a first type with unchanged attributes, a second cost is associated with preserving data of a second type with one or more changed attributes, and the first cost and the second cost are set to discourage preserving data of the second type more than preserving the data of the first type.

19. The method of claim 1, wherein tree data is at least partly from the first tree.

20. The method of claim 1, wherein tree data is at least partly from the second tree.

21. The method of claim 1, wherein the second tree is received if the second tree is different from the first tree.

22. The method of claim 1, further comprising:

if two or more corresponding data are found, then:

selecting larger selected data, at least part of the larger selected data including a larger subtree in a first tree representation of the first set of data, the larger subtree including the selected data;

determining a second edit sequence between at least part of the first set of data and at least part of a second tree representation of the second set of data, the first set of data including at least part of the larger selected data, the second edit sequence including any of insertions, deletions, and substitutions;

finding corresponding data of the second set of data, the corresponding data having a correspondence to the larger selected data, the correspondence at least partly found by determining the second edit sequence; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the second edit sequence.

23. The method of claim 1, wherein the correspondence is at least partly found by one or more of: determining the edit sequence, at least part of at least one of a first plurality of paths from a root of a tree representation of the first set of data to selected data of the tree representation of the first set of data, at least part of at least one of a second plurality of paths from a root of a tree representation of the second set of data to corresponding data of the tree representation of the second set of data, and one or more edit sequences between at least one of the first plurality of paths and at least one of the second plurality of paths.

24. The method of claim 1, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

25. The method of claim 1, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

26. The method of claim 1, wherein the first tree and the second tree represent different trees.

27. The method of claim 1, wherein the first tree and the second tree represent a same tree.

28. The method of claim 1, wherein the first tree and the second tree represent different versions of a same tree.

29. The method of claim 1, further comprising:

determining at least one edit sequence of forward and backward edit sequences between at least part of the first tree and at least part of the second tree;

performing at least one of 1) and 2):

1a) pruning a relevant subtree from at least part of the first tree, the relevant subtree at least partly determined from the forward and backward edit sequences;

1b) determining a pruned edit sequence between the pruned relevant subtree and at least part of the second tree;

2a) pruning a relevant subtree from at least part of the second tree, the relevant subtree at least partly determined from the forward and backward edit sequences;

2b) determining a pruned edit sequence between at least part of the first tree and the pruned relevant subtree; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the pruned edit sequence.

30. A method of extracting relevant data, comprising:

accessing at least a first set of data of a first tree, wherein the first set of data includes selected data of the first tree, the selected data at least partly specifying tree data;

accessing at least a second set of data of a second tree;

determining a second path from a root of the second tree that corresponds to a first path from a root of the first tree to the selected data; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly determined by the second path from the root of the second tree.

31. The method of claim 30, wherein the second path is determined at least in part by:

traversing the first tree and the second tree;

at each traversed level of the first tree, the traversed level of the first tree including a plurality of level nodes, selecting a level node of the plurality of level nodes, the level node being in the first path;

at each traversed level of the second tree, selecting a best corresponding node at the traversed level of the second tree, the best corresponding node saving a best correspondence to the selected level node of the plurality of level nodes.

32. The method of claim 31, wherein the best corresponding node is determined at least in part by determining an edit sequence between a first subset of data obtained from at least part of the first set of data and a second subset of data obtained from at least part of the second set of data, the edit sequence including any of insertions, deletions, substitutions, matches, and repetitions.

33. The method of claim 32, wherein the repetitions are subtrees of at least the first tree.

34. The method of claim 32, wherein the edit sequence includes at least two repetitions, the at least two repetitions based on at least one subtree of the first tree, and the at least two repetitions appears in the second tree, and the at least two repetitions include at least a first subtree and a second subtree, and the first subtree has at least one difference from the second subtree.

35. The method of claim 34, wherein each of the at least two repetitions is obtainable from the at least one subtree of the first tree by some sequence of one or more insertions, deletions, substitutions and matches.

36. The method of claim 32, wherein the edit sequence includes none of insertions, deletions, substitutions, matches, and repetitions.

37. The method of claim 32, wherein the edit sequence includes at least one of one or more insertions of nodes, one or more insertions of subtrees, one or more deletions of subtrees, one or more deletions of subtrees, one or more substitutions of nodes, one or more substitutions of subtrees, one or more repetitions of nodes, and one or more repetitions of subtrees.

38. The method of claim 32, wherein the edit sequence is at least partly determined by calculating a total cost, and each of one or more of insertions, deletions, substitutions, and matches is associated with one or more costs.

39. The method of claim 38, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more matches between at least some selected data of the first tree and at least some data from the second tree.

40. The method of claim 38, wherein the one or more costs are at least partly set to encourage the edit sequence to include one or more repetitions.

41. The method of claim 38, wherein a first cost is associated with a first match at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second match at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

42. The method of claim 38, wherein a first cost is associated with a first insertion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second insertion at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

43. The method of claim 38, wherein a first cost is associated with a first deletion at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second deletion at a second distance from a root of a

tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

44. The method of claim 38, wherein a first cost is associated with a first substitution at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second substitution at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

45. The method of claim 38, wherein a first cost is associated with a first repetition at a first distance from a root of a tree representation of some set of data, a second cost is associated with a second repetition at a second distance from a root of a tree representation of some set of data, the first distance is less than the second distance, and the first cost and the second cost are different.

46. The method of claim 38, wherein a first cost is associated with a first text-based content substitution such that a first length of substituting text-based content is substantially equal to a first length of substituted text-based content, a second cost is associated with a second text-based content substitution such that a second length of substituting text-based content is substantially different from a second length of substituted text-based content, and the first cost and the second cost are set to discourage the second text-based content substitution more than the first text-based content substitution.

47. The method of claim 38, wherein data includes at least a first type and a second type, and the one or more costs are at least partly set to discourage substitutions of one or more of the first type for one or more of the second type.

48. The method of claim 38, wherein data includes at least a first type and a second type, and the one or more costs are at least partly set to discourage substitutions of one or more of the second type for one or more of the first type.

49. The method of claim 38, wherein a first cost is associated with preserving data of a first type with unchanged attributes, a second cost is associated with preserving data of a second type with one or more changed attributes, and the first

cost and the second cost are set to discourage preserving data of the second type more than preserving the data of the first type.

50. The method of claim 32, wherein the first subset of data includes nodes in the first tree that are within a first neighborhood of any of the selected level nodes of the traversed levels of the first tree, the selected level nodes of the traversed levels of the first tree being on the first path from the root of the first tree to the selected data, and the second subset of data includes nodes in the second tree that are within a second neighborhood of children nodes of the best corresponding node selected at a previous level in the second tree.

51. The method of claim 50, wherein the first neighborhood of any selected level node includes a first plurality of close nodes according to a first distance measure, and the second neighborhood of a child node includes a second plurality of close nodes according to a second distance measure.

52. The method of 51, wherein the first distance measure between any selected level node and another node is at least partly determined a first number of tree edges between any selected level node and another node the second distance measure between the child node and another node is at least partly determined a second number of tree edges between the child node and another node.

53. The method of 51, wherein the first distance measure between any selected level node and another node is at least partly determined a first number of tree levels between any selected level node and another node the second distance measure between the child node and another node is at least partly determined a second number of tree levels between the child node and another node.

54. The method of claim 30, wherein one or more of the first set of data and the second set of data is represented at least partly by a tree.

55. The method of claim 30, wherein one or more of the first set of data and the second set of data is represented at least partly by a set of linearized tokens.

56. The method of claim 30, wherein the first tree and the second tree represent different trees.

57. The method of claim 30, wherein the first tree and the second tree represent a same tree.

58. The method of claim 30, wherein the first tree and the second tree represent different versions of a same tree.

59. The method of claim 30, wherein tree data is at least partly from the first tree.

60. The method of claim 30, wherein tree data is at least partly from the second tree.

61. The method of claim 30, wherein the second tree is received if the second tree is different from the first tree.

62. An apparatus for extracting relevant data, comprising:

a plurality of one or more computing devices adapted to perform:

accessing at least a first set of data of a first tree, wherein the first set of data includes selected data of the first tree, the selected data at least partly specifying tree data;

accessing at least a second set of data of a second tree;

determining an edit sequence between at least part of the first set of data and at least part of the second set of data, the edit sequence including any of insertions, deletions, substitutions, matches, and repetitions; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly found by determining the edit sequence.

63. An apparatus for extracting relevant data, comprising:

a plurality of one or more computing devices adapted to perform:

accessing at least a first set of data of a first tree, wherein the first set of data includes selected data of the first tree, the selected data at least partly specifying

tree data;

accessing at least a second set of data of a second tree;

determining a second path from a root of the second tree that corresponds to a first path from a root of the first tree to the selected data; and

finding corresponding data of the second set of data, the corresponding data having a correspondence to the selected data, the correspondence at least partly determined by the second path from the root of the second tree.